# Preserving digital heritage: a network centric approach

Ana Rodrigues
DGLAB*
Alameda da Universidade
1649-010 Lisboa
ana.rodrigues@dglab.gov.pt

Francisco Barbedo
DGLAB*
Alameda da Universidade
1649-010 Lisboa
francisco.barbedo@dglab.gov.pt

Lucília Runa
DGLAB*
Alameda da Universidade
1649-010 Lisboa
lucilia.runa@dglab.gov.pt

Mário Sant'Ana
DGLAB*
Alameda da Universidade
1649-010 Lisboa
mario.santana@dglab.gov.pt

## ABSTRACT

The paper presents a rationale about the project digital continuity launched by DGLAB, explaining it's fundaments, methodology and findings so far. It finishes proposing future work and aims to be achieved.

## Keywords

Digital heritage, cooperation network, digital preservation

## 1. INTRODUCTION

Preserving digital objects is no longer an exclusive technological challenge. Correlated with informatics development and use, social and organizational issues are mandatory in order to obtain complete and accurate preservation solutions.

Three orders of reasons contribute to this situation: the first one is that digital preservation is a pervasive problem that spreads to every organization and individual that produce professionally or individually digital data.

The second one is that lack of preservation actions drive very quickly to obsolescence, which is a condition that can actually stop business continuity. Today's organizations are beginning to grasp this reality as the digital data that has being produced for the past years accumulated into a proportion to which digital obsolescence is already being strongly perceived..

The third reason is that preserving digital objects is a costly activity that demands a lot of expertise and highly qualified people but also a considerable investment in equipment and development as well as a high fixed cost in order to keep digital repositories. This reality presently highlighted by financial crash and general economic depression may lead to establishing partnership between institutions that need to preserve digital material sharing costs and knowledge- A powerful and dedicates info structure and human capacitation on order to deal effectively with it it's a business that is better managed together- Preserving digital objects is a solidarity activity and not an egotistic one.

In this process the difference between cultural domains seems to become less relevant as critical mass is best achieved converging efforts from digital heritage holders irrespectively of the community of practice (CP) to which each one belongs to.

Working together means sharing resources and to manage collectively organizational data. This requires a new way of doing business, which in turns demands new social relationships, new management models and new financial sustainability solutions.

This problem concerns every digital object that must be preserved for operational or cultural reasons more than 7 years. But as far as digital heritage is concerned the issue is permanent and requires particular attention as heritage belongs to every citizen in a nation.

Portuguese National Archives decided to organize a meeting in order to listen to different stakeholders regarding the problem of preserving digital heritage and explore the possibility of organizing new ways, network centered, of preserving digital heritage.

The event took place in September 2013 and the agenda consisted on putting together different CP from public or private sector in order to discuss the problems each experienced in preserving digital objects.

Some questions were raised concerning the possible convergence of similar problems independently of the cultural domain to which digital material belongs. We also wanted to harvest people perception on what is really digital heritage, meaning to know how different CP included digital objects into their cultural and heritage domain. We were particularly interested in digital surrogates of analogical digital material that have been created and managed through massive digitization projects that were in the past decade very popular. Should digital surrogates be processed as originals in the sense that they should be preserved forever?

Other questions were asked and also raised by the meeting attendants, such as the convergence between different CP when dealing with digital data; the possibility of shared solutions of storage and digital repositories or issues related to cloud solutions.

The conclusions [1],were drawn by the interventions of invited speakers together with the conclusions of 4 workshops held with the public that were organized of 4 thematic issues: 1/ The inclusion of digital objects into the heritage set; 2/

[1] http://1seminariopreservacaopatrimoniodigital.dglab.gov.pt/conclusoes/

Curatorial responsibilities in digital world; 3/ Common technological platform; 4/ Building a national network for preserving digital heritage.

# 2. THE PROJECT

The project, baptized as "digital continuity", is an initiative of National Archives and was launched as a response to the conclusions from the 1st seminar that considered important to continue the work started and to raise awareness on the present situation in Portugal regarding preservation of digital heritage and to propose network centered solutions to that issue.

The project assumes that in digital environment, heritage objects, no matter their cultural provenance, are basically information binary coded and machine readable.

This fact turns digital information as digital heritage sharing common features that may enable their common management.

The differences between different CP objects become indistinct, except for the use that specific communities of audience require. But even then recent organizational experiences like Europeana show us that the needs of remote audience does not account for the traditional division between different cultural domains.

The basic organization of the project staff is to join together representatives of different cultural domains and CP in order to discuss the possibility of constructing such a structure looking for similarities and differences spread by multiple layers of practice and knowledge: the topics we want to discover are:

1/ regulatory framework comprehending law, terminology structures and concepts, metadata standards and formats;

2/ authenticity and appraisal considered under the different point s of view of the represented CP;

3/ access requirements and Digital Rights Management (DRM) in digital landscape;

4/ technical requirements such as storage, dimension, prospective growth;

5/ architecture and logical model definition;

6/ business and sustainability models for the network.

The project is organized on a bottom up methodology that stems in grounded theory, enabling new findings to lead to new analysis trends. The approach is included in an inductive type and is inspired on international experiences such as project Interpares[2] and NDIIPP[3].

The project development plan that will eventually lead to a pack of conclusions that cannot, for the moment, be anticipated. The ongoing work will inform the actions to be taken in the future. It is possible to achieve a situation where different levels of acceptance from the participants are identified.

The project began in January 2014 and has two phases: the first one that is currently going on aims to produce a body of knowledge aligned with different layers of research that may inform the participants of the advantages and disadvantages or a common preservation info structure. We also intend to develop a governance and sustainability model that enables a smooth management and operation of a common network. Acquiring financial resources to bid whatever is decided at the end of phase one will also be a subject that the working team, will tackle by 2015.

Several people that work in different CP were invited to join the team, which was divided in an executive set of people committed to gather material, ensure the logistics and perform analysis and another set whose task is basically to provide data and information and also discuss and validate analyzed data submitted to them.

All the work developed so far has considered all the CP that the project team members belong to:

- Archives;
- Libraries;
- Cultural Heritage;
- Journalism;
- Television;
- Radio;
- Cinema;
- Photography;
- Music;
- Multimedia, entertainment.

# 3. WORK DEVELOPED SO FAR

## 3.1 Step one

The first step of the planned chronogram considered harvesting the regulatory environment that influences the activity of the different CP represented in the project team.

This phase is already finished and led to the general conclusion that there are more similarities than differences between requirements and factors that influence the work of CP. This observation corroborates the work already developed in international instances, like Europeana or NDIIP[4].

In order to preserve together digital material there it is only necessary a common set of requirements and practices that enable trough easy interoperability the development and operation of common structures.

The methodology followed was to harvest systematically the documents on the defined sets of observation. This was achieved with the help of project team that provided all information regarding their specific cultural domain. The analysis then took place on a sample of all the documents identified This sampling was justified by 1/ as some of them are not really in use by CP, at least in Portugal and 2/ identified documents are repetitive not bringing new data to the analysis. For example there are several standards and terminologies on music or photography that partially or fully overlap. In this case it would be useless to take all of them in account for the analysis performance.

A comparison was then performed element by element in order to find similarities and divergences between them. All the results were intensively discussed with the project team.

### *3.1.1 Law and regulations*

In the context of legislative analysis, for each CP, we proceeded to the recognition and identification of regulatory statutes governing the respective activity - legal regimes Act, deontological codes - as well as specific aspects with

[2] http://www.interpares.org/

[3] http://www.digitalpreservation.gov/index.php

[4] http://www.europeana.eu/; http://www.digitalpreservation.gov/

particular relevance to digital preservation, which were grouped around these two categories.

Accordingly, we identified the regulatory and specific law - both national and European level, this latter whenever possible[5].

From the inquiry conducted, it was found the existence of acts:

- With interest for all CP (Law on Copyright and Related Rights, the Legal Deposit Act and Data Protection Law);
- Multidomain applicable to journalistic activity - exercised through the Press, Radio and Television - and Cultural Heritage;
- With explicit references to heritage preservation / digital, including:
  - Clause 11 of the Law on Cultural Heritage on "duty of preservation, protection and enhancement of Cultural Heritage";
  - Chapter VII of the Television Act, paragraph 1, 2 and 3 of article 92 on "preservation Television Heritage";
  - Chapter VII of the Radio Law, article 83 on "Heritage preservation radio broadcasting - records of public interest".[6]

### *3.1.2 Terminology*

Regarding terminology and vocabulary structures were studied. It was observed that 2 types of vocabulary structures may be find. A first one dedicated to the activity itself. It contains concepts and vocabularies whose purpose is to help workers that deal with that specific core business. For example movie terminology corresponds to vocabulary that is actually used in the cinematography industry and therefore contains terms and concepts dedicated to that particular business.

The other type consists in vocabulary structures dedicated to the activity of describing or cataloguing material that was produced for a specific kind of activity. Photography for instance shares a set of common terms to other structures that deal e.g. to description works of art.

Although there are of course a lot of different concepts and therefore terms, it is possible to find a core of common terms and concepts that indicate an approach between the way different CP conceptualize aspects that are common to their activities and the material they deal with.

The presence of the same terms (that represent concepts) in several vocabulary structures may give a clear indication of the existence of a shared perception of at least some parts of different CP. The analysis of the 6 vocabulary structures [7]had as an outcome the identification of a set of terms that exist at least in more than two of the examples observed. The average number of common terms present in the observed vocabulary structures is 4,2.

The results by CP and more popular terms are depicted in the following figure and table.

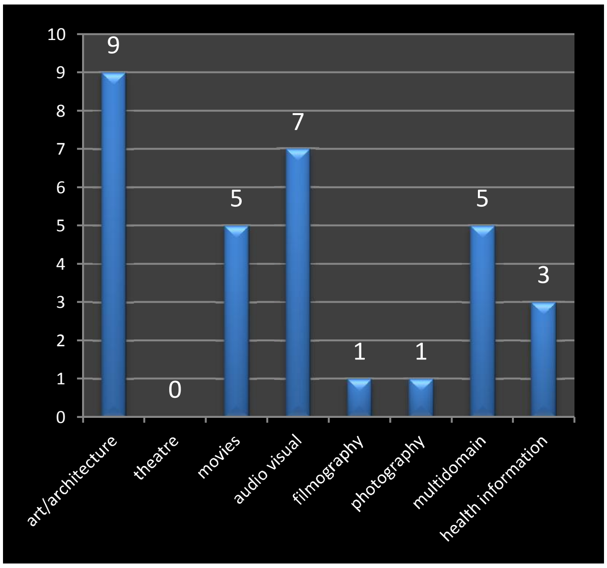

**Figure 1. Popular terms by CP**

**Table 1. More popular terms by CP**

| **terms** | **# occurrences** |
|---|---|
| Access | 5 |
| Authenticity | 2 |
| Appraisal | 4 |
| Custody | 1 |
| Identification digital heritage | 1 |
| Digital heritage | 4 |
| Digital preservation | 3 |
| Certification and security of repositories | 2 |
| Copyright | 5 |
| Usability | 2 |

### *3.1.3 Standards*

The identified standards [8]correspond to:

[5] Regarding methodology followed, see: http://1seminariopreservacaopatrimoniodigital.dglab.gov.pt/projeto-continuidade-digital/documentos-de-projeto/.

[6] In spite of the two last explicit references, there are no public laws to protect or to classify the television heritage but only an organizational policy management of collection.

[7] Other vocabulary structures were identified. Those can be checked in 1st project report available at: http://1seminariopreservacaopatrimoniodigital.dglab.gov.pt/wp-content/uploads/sites/19/2014/10/SinteseMaterialPassoUm1.0.xlsx.

[8] Listed in the tab “Standards”, *Annex 1* of the 1st project report available at: http://1seminariopreservacaopatrimoniodigital.dglab.gov.pt/wp-content/uploads/sites/19/2014/10/SinteseMaterialPassoUm1.0.xlsx.

- International rules, standards, guidelines and recommendations prepared by International Council on Archives (ICA), International Federation of Television Archives (IFTA), International Council of Museums (ICM), International Federation of Library Associations and institutions (IFLA), Music Library Association (MLA), Online Audiovisual Catalogers (OLAC), International Organization for Standardization (ISO), Europeana, Archives Portal Europe Network of Excellence (APEx), etc.
- National rules, prepared by national archives, libraries and museums (Portugal, Canada, United States of America, Australia).

Their distribution is as follows:

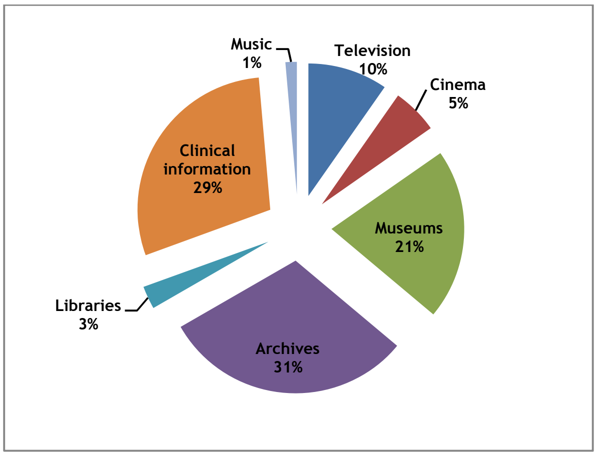

**Figure 2. Standards distribution**

From the identified standards only 4 are shared by two CP:

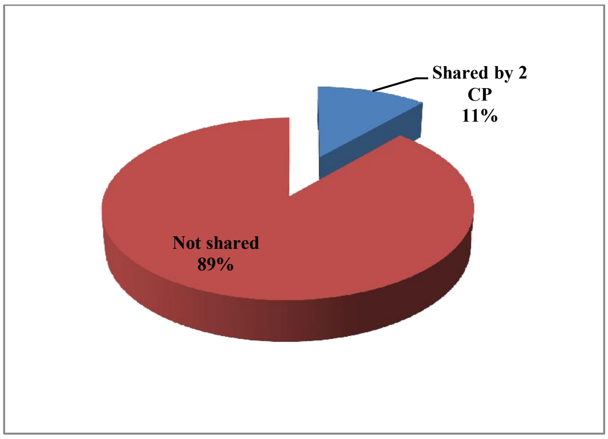

**Figure 3. Shared standards**

They were classified according their pertinence to the main goal of the working team: represent/describe cultural heritage objects.

The purpose was to proceed to a comparative analysis of the most pertinent standards, detect possible matches and to ascertain the possibility of mapping their elements to a common structure.

Considering the clinical information and the specificity pointed to the clinical records, several standards were identified. Case studies were conducted to ascertain their correspondence with standards used by other CP.

The main conclusions of the standard analysis were as follows:

1/ the major part concern to a unique CP. The exceptions are CIDOC-CRM and Dublin Core;

2/ the major part are intended only for one of two things: objects description or objects description encoding. The exceptions are Dublin Core, CIDOC-CRM, EBU-TECH 3293, MPEG 7;

3/ the major part focus on the object. An exception to CIDOC-CRM, event-centred and object-oriented;

4/ the major part include contextual information about the described objects, although archives and museums are the CP who considers this kind of information with a greater depth;

5/ there are other CP, besides the archives, which adopt a multilevel description, like museums and libraries;

6/ the major part of the standards are categorial, which means they group the information in areas and, within these areas, in different elements. The exceptions are Dublin Core, CIDOC-CRM, EBU-TECH 3293 (based on Dublin Core) and MPEG 7, which are combinatorial: they use metadata, metadata schemes and description definition languages;

7/ all the standards assume a concern about cross-domain and multidomain coordination. This concern, in the case of the archives, is reflected, e.g., in ISAAR (CPF): "(…) separate capture and maintenance (…) of contextual information is a vital component of archival description. The practice enables the linking of descriptions of records creators and contextual information to descriptions of records from the same creator(s) that may be held by more than one repository and to descriptions of other resources such as library and museum materials that relate to the entity in question. Such links improve records management practices and facilitate research." (ISAAR (CPF), 1.5, p. 7);

8/ standards have a common goal: sharing and retrieval information;

9/ categorial rules are more classic and the first to be produced, but they are quite equivalent. Considering their specificities:

- some of them are related with the scope of the standards: e.g. ISAD (G) is a general archival standard – "(…) rules (…) do not give guidance on the description of special materials such as seals, sound recordings, or maps." (1.4, p. 7). The same principle does not apply to libraries: the ISBD includes specific rules to specific materials;
- but there are other specificities which are not related with the characteristics of the described objects: e.g. elements accommodating information about objects location are considered only by the museum standards, although this kind of information is also relevant to libraries and archives;
- there are also elements intended to accommodate equivalent information. However, considering their specific objects, there are content specificities: e.g. "Immediate source of acquisition or transfer" (archives), intended to "record the source from which the unit of description was acquired and the date and/or method of acquisition" and "Acquisition method" (museums),

versus “Acquisition modality” (libraries), intended to accommodate information like the price of a resource;

- on the contrary, there are elements specially intended to specific types of objects: e.g. the ones grouped in the “Edition area” or in the “Publication, production, distribution area” - Libraries) -, the last one having correspondence in the standards used by CP like television.

10/ combinatorial rules, by their relational approach, offer a flexible description compatible with different CP;

11/ Practical experiences of crossed description proved the possibility to describe a specific domain object using the rules of another domain.

### *3.1.4 Formats*

The central purpose of the activity was to identify the formats mainly used among the different CP.

All formats were considered, whether they are used with the objective of access or preservation.

In order to be referenced by all CP, an open formats list in an excel file was drafted. Each CP could add any formats they considered adequate. The list was organized by categories in order to allow conceptual framework according to the website of the US Library of Congress “The Digital Formats Web site”[9] and several *ad hoc* contributions of the digital continuity project team.

The list contained 176 formats structured for 6 categories (still images, sound, text, generic, moving images and datasets) of the 8 provided (Geospatial and Web Files were not considered).

In this section (3.1.4), the Photography and Radio CP, not present on the project team, answered the survey.

Results (some of the more salient aspects):

- A total of 54 different formats are used by all surveyed CP[10].
- The categories with a higher number of formats are Sound (47) and Still Images (42). The categories with fewer representative elements are Text (16) and Generic (13). With numbers between these two we can find Moving Images (38) and Datasets (20).
- No format is used by all CP in the study
- Moving images, with 23 hits, is the category that has the highest number of formats in use. In the other end, Datasets, with 4 hits, is the category less mentioned

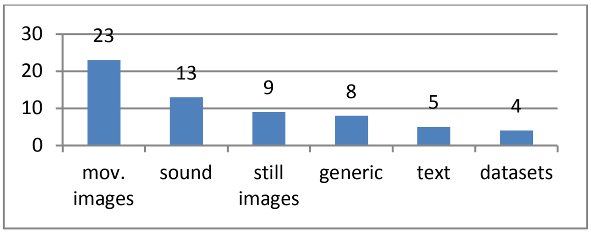

**Figure 4. Number of formats referred more often by category**

- Formats used by a greater number of CP

XML (identified in the Generic category), PDF and JPEG are the most used formats by the different CP. The MP3 and TIFF formats are referred by 5 CP, while the PNG, PPT and ZIP are mentioned by 4 CP.

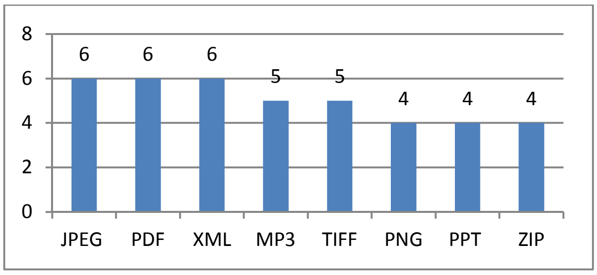

**Figure 5. Formats used by a greater number of CP**

- Most used format by category

XML for generic, PDF for text and JPEG for still images are used by 6 different CP. MP3 for sound is used in 5 CP. AVI, MPEG-4 Video Encodings and Quicktime for moving images are used by 3 CP. XLS for dataset is used in 2 different CP.

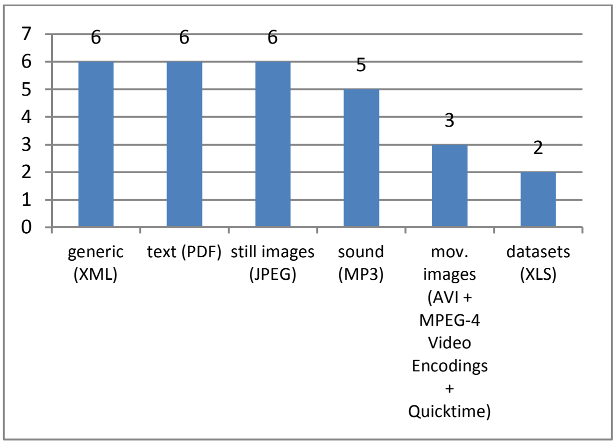

**Figure 6. The most used formats by category**

- Formats used by the CP by category (summary)
  - Still images has 9 formats: JPEG, TIFF, PNG, BigTIFF, DICOM, DNG, JPEG 2000 Encodings, JPEG 2000 File Formats and GIF;
  - Sound has 13 formats: MP3, WAVE, QuickTime, WAV, AIFF, ASF, BWF, FLAC, ID3, MP2, PCM, WM (Windows Media) and Music XML;
  - Text has 5 formats: PDF, DOC, DOCX, XML and TXT;
  - Generic has 8 formats: XML, PPT, ZIP, AUTOCAD, ASF, DGW, ISO_image and RIFF;
  - Moving images has 23: AVI, MPEG-4 Video Encodings, QuickTime, Flash (SWF, FLA, FLV),

[9] Cf. *Sustainability of Digital Formats Planning for Library of Congress Collections* [Online]. [Accessed on, October 2014] at WWW <URL: http://www.digitalpreservation.gov/formats/index.shtml>.

[10] In this point, formats marked in more than one category, such as ASF (marked as Sound, Generic and Moving images) were considered only once

WM (Windows Media), DivX, MPEG-2, AAF, AC-3, ASF, Cinepak, DPX, DTS, DV, H.26n ITU-T video encoding standards, Indeo, MPEG-1, MPEG-4 File Formats with Encoded Bitstreams, MXF, RealMedia, Uncompressed video encodings, Digital Betacam, AKA Digibeta or D-Beta, HDCAM;

- Datasets has 4 formats: XLS, CSV, ISO_8211, XLS (linux).

## 3.2 Step 2

The second step aimed to identify the levels perceived of the different CP on two aspects that are crucial:

- The perception of authenticity requirements that impend over digital objects from different cultural domains;
- The methodology and criteria, if any, used by CP in order to identify and classify digital objects as "worthy" of becoming part of digital heritage.

Both aspects can have a dramatic impact on a future common repository. It is therefore necessary to be aware of different realities concerning these issues.

The methodology used for this task was an enquiry to which the working team was asked to provide answer. Although the small number of respondents could not give definitive or meaningful answers the large representative basis of different CP might give an accurate perception of sensibilities and trends on the topics.

The enquiry served a second purpose of testing and allowing adjustments to the questions asked that will be used in the construction of a second big survey that the project intends to launch to nearly 300 cultural, both public and private, portuguese cultural institutions.

The results were processed with descriptive statistical measures: the mean in order to measure the degree of agreement (central tendency) of opinions expressed; the standard deviation, as a dispersion measure, used to confirm the convergence depicted by centrality measures. The SD would depict the level of "disagreement" to the general trend estimated by the mean.

The most relevant conclusions are as follows:

1/ usually people belong to more than one CP. In fact there is a fairly common situation where an institution has custody of objects from several cultural domains, processed accordingly to the CP that usually has expertise on them. E.g. a museum may hold archival collections that will be processed in a compliant way to archival standards, while at the same time, as far as museum objects are concerned, other specific standards will be adopted;

2/ globally the results revealed a remarkable convergence on the 7 variables under scrutiny corresponding to authenticity. There were some exceptions essentially due to some lack of practice and theoretical thought on the subject. This was particularly noticed in CP journalism, where there is a poor tradition on heritage connected issues.

3/ the divergences concern some digital object's authenticity requirements from different cultural domains were not meaningful being reported an average of 4,52 on a 1 to 5 scale.

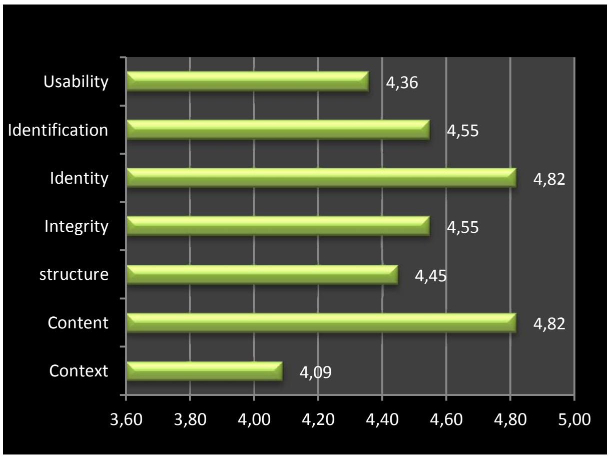

**Figure 7. Relevance evaluation to authenticity features**

4/ it was observed a fair distance between CP regarding criteria for evaluating objects as deserving integration into heritage set. Museums, archives and libraries adopt more or less the same number of criteria, while other CP, such as music, adopt less criteria. No real conclusion can be inferred from this observation because everyone, except journalism, has actually criteria for evaluating digital heritage;

5/ there seems to be some lack of formal criteria to evaluate objects as heritage to be. Exception made for archives, which have a very regulated evaluation process, the most common situation is included in the area of collection management which from each institution policy and strategy. As such it may vary between organizations and even inside the same CP;

6/ several exclusion factors were mentioned, which is interesting to remark as people consider different factors, usually absent on traditional environment, that may influence decisively the classification of objects as heritage;

7/ costs regarding classification, access and storage were mentioned as possible hindrances for preserving digital objects as heritage.

## 4. CONCLUSIONS SO FAR AND FUTURE ACTIONS

The analysis developed along the first 2 steps leads to the conclusion that no big issues separate different CP represented in project team. And until now there could not be found any major issues that might put insurmountable obstacles on the building of a common network to share resources and preserve common digital heritage.

Financial and technical aspects must be clarified prior to any stakeholder to make a decision about it's possible participation on such a network.

Next action, as predicted in he approved chronogram, is launching of a big survey that hopefully will bring us data that will validate or contradict the preliminary observations harvested by debate and enquiries inside the project team.

The development of a study regarding business model of the prospective network as well as the possible financial models and sustainability will be the most difficult challenges of this project. We expect to gather support of experts from different areas such as economics, sociology and engineering, maybe connected to academic world to cooperate with the project team and tackle these particular issues on a knowledgeable way.

According to the project schedule, next tasks aim to complete the steps of the first phase of the project - in principle until February 2015. Data will be gathered relating to:

- physical environment: storage technology platforms (information size and growth estimates);
- logical environment (software);
- conceptual and logic architecture (network model, including the model of governance and sustainability).

(*) Directorate General of Book, Archives and Libraries